Power analysis for efficacy RCT based on pilot safety RCT

Quentin J.M. Huys

When doing power analyses, we try to estimate how large the sample has to be so that, after running the study, we have a certain degree of certainty that if an effect of a certain size was there we didn't miss it. The effect size the size of the difference between the groups, and does not refer to whether this size of difference is significant. An effect of size d may be visible in a small study, but not statistically significant. So significance and effect size are different measures.

When planning a study, we ask how large the sample needs to be so that for an effect of a given size we are likely to observe a significant effect with some probability. This probability is the power. It is standard for efficacy trials to aim for a power of 90%, i.e. for a probability of 90% that a group comparison will be significant, given that in truth there was a group difference of a certain assumed size d.

The effect size and hence power estimates obviously depend a lot on whether we do an ITT or PP analysis: they are much larger in the PP than the ITT analyses, because of course ITT includes people who did not do the therapy really.

We ran a safety RCT comparing waitlist to the intervention component of the app alone. 50 participants were randomized to each group. We'd now like to estimate the size of the new efficacy RCT based on these results. The results are shown in figure 1. The top row shows the curves for all participants (ITT, left), only those who did the therapy (PP, middle), and then also the analysis when subtracting baseline SPIN scores in the PP analysis. The effects are weakest in ITT, as one would expect. In this sample, the effects are stronger when looking at change scores (right).

The bottom row of plots in figure 1 shows the power for different sample sizes, when basing the effect size on the respective definitions/analyses. The dashed red line shows the 90% power, so that determines the sample size we should shoot for.

As we want to make statements about the app as a whole, it may make sense for the primary analysis to be on an ITT sample. This would measure overall improvement in the two groups independent of whether participants actually engaged in the app itself. But it means that the result is very meaningful as it also captures how much people do want to engage with the app, which will e important for how well it does in reality.

Judging from the pilot RCT, the ITT effect size at week 4 is between 0.2 and 0.34 (see top left panel). Basic power analyses and simulations suggest that we would need a sample of around 250 for this. I would suggest this is what we aim for.

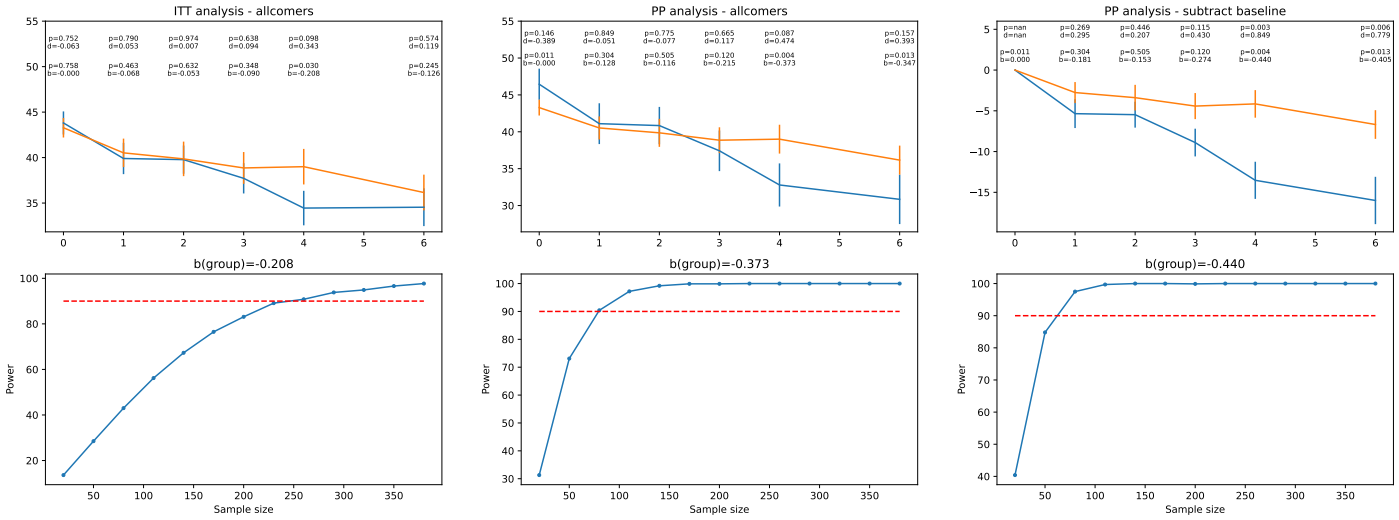

FIGURE 1: Analysis of RCT data. Orange is waitlist, blue intervention. The top row shows SPIN scores and analyses, the bottom row power analyses. **Top left**: ITT analysis. The top row of statistics shows simple t-tests and Cohen's d for each week. The bottom row shows standardized β estimates for the group effect for a linear regression with baseline spin, age and group, with the dependent variable being the spin score at each of the weeks. **Bottom left**: Power analysis for ITT analysis. I extracted the week 4 beta values from the regression (bottom line of results in the panel above), and simulated this, counting the number of times a significant result was achieved. A simple power analysis for a 2-sample t-test for a Cohen's d of 0.343 yiels an $N = 294$. Repeating this but using as effect size the standardized regression coefficient of 0.208 yields an $N = 794$. The simulate power analysis suggests a total sample size of $N = 250$. **Top middle**: PP analysis. This only includes participants who were marked as 'completers'. The top row of statistics again shows simple t-tests and Cohen's d for each week. The bottom row shows standardized β estimates for the group effect for a linear regression with baseline spin, age and group, with the dependent variable being the spin score at each of the weeks. **Bottom middle**: Power analysis for PP analysis. I extracted the week 4 beta values from the regression (bottom line of results in the panel above), and simulated this, counting the number of times a significant result was achieved. This suggests a total sample size of $N = 90$. A simple power analysis for a 2-sample t-test for a Cohen's d of 0.474 yields an $N = 154$. Repeating this but using as effect size the standardized regression coefficient of 0.373 yields an $N = 248$. **Top right**: PP analysis on difference scores. This only includes participants who were marked as 'completers'. Baseline values were subtracted from subsequent values, so this is an analysis of difference scores from baseline. The top row of statistics again shows simple t-tests and Cohen's d for each week. The bottom row shows standardized β estimates for the group effect for a linear regression with baseline spin, age and group, with the dependent variable being the spin score at each of the weeks. **Bottom right**: Power analysis for PP analysis on difference scores. I extracted the week 4 beta values from the regression (bottom line of results in the panel above), and simulated this, counting the number of times a significant result was achieved. This suggests a total sample size of $N = 76$. A simple power analysis for a 2-sample t-test for a Cohen's d of 0.849 yields an $N = 50$. Repeating this but using as effect size the standardized regression coefficient of 0.440 yields an $N = 180$.